Brustkrebs Indicatoren

Timo Michaelis

3. Januar 2025

Zusammenfassung

Im Folgenden wird versucht anhand verschiedener Indikatoren zur prognostizieren, ob eine Patientin Brustkrebs besitzt. Dieses Model dient lediglich der Hilfestellung, nicht aber dem Ersatz eines fachkundigen Arztes

Inhaltsverzeichnis

1 Einleitung **2**

2 Datensatz **2**

3 Methoden **6**
3.1 Random Forest Classifier 6
3.2 Support Vector Machine 8
3.2.1 Kernel 8

4 Ergebnisse **8**
4.1 Random Forest Classifier 8
4.2 Support Vector Machine 11

5 Diskussion **14**

6 Fazit **16**

A Anhang: Erklärung **17**

B Anhang: Jupyter Notebook **17**

1 Einleitung

In der vorliegenden Arbeit wird eine Unterstützung zur Erkennung von Brustkrebs geboten. Dafür werden verschiedene Eigenschaft u.a. *durchschnittlicher Radius, durchschnittliche Textur, usw.* berücksichtigt und eine wahrscheinliche Prognose geboten.
Das Ziel dieser Arbeit ist es somit ein Modell zu erstellen, welches bei Input der benötigten Parameter eine vorläufige Diagnose stellt, welche etwaig dem Patienten bzw. dem Arzt zu einer genaueren Untersuchen verleiten. Dies verringert somit nicht nur die die Chance von sogenannten *false positives*, sondern kann auch die Bereitschaft unterstützen häufiger notwendige Untersuchungen durchzuführen, da die Untersuchung weniger komplex wird.
Im folgenden wird grundsätzlich der Ursprung und Inhalt des Datensatzes diskutiert, sowie die darauf angewandten Methoden näher erläutert.

2 Datensatz

Der verwendete Datensatz stammt aus der *Diagnostic Wisconsin Breast Cancer Database.*$\text{breast}_c ancer_w isconsin_d iagnostic_1 7, wieerwartetbefasstsichdieserDatensatzmitGe$

mean radius

mean texture

mean perimeter

mean area

mean smoothness

mean compactness

mean concavity

mean concave points

mean symmetry

mean fractal dimension

radius error

texture error

perimeter error

area error

smoothness error

compactness error

concavity error

concave points error

symmetry error

fractal dimension error

worst radius

worst texture

worst perimeter

worst area

worst smoothness

worst compactness

worst concavity

worst concave points

worst symmetry

worst fractal dimension

Hinzu kommt noch der *Target-Datensatz*, welcher beschreibt ob der Datenpunkt jeweils *Malignand* oder *Benign* ist.
Bevor näher auf den Datensatz und seine zusammenhänge eingegangen wird, gilt es zu beurteilen inwieweit dieser aufbereitet ist. Hiefür sei relevant zu testen ob der Datensatz vollständig und inherrent logisch ist, sollte dies nicht der Fall sein, so gilt es diesen zu bereinigen, durch etwaiges löschen bzw. ersetzen.
Insofern wurde u.a. überprüft ob es null oder negative Werte gibt, solche dürfte es nämlich nicht geben. Der Datensatz ist frei von solchen Werten, als letzes .
Abseits der Integrität dieses Datensatzes sei noch dessen Koheränz zu überprüfen. Dafür wurden mehrere Boxplot-Graphen12 erstellt, welche wie ersichtlich Ausreißer angeben, 74 sind insgesamt mit 3σ Regel zu finden. Nun gilt es die

Frage zu stellen, ob diese Ausreißer entfernt werden müssen.
Letzlich zielt diese Frage darauf ab, ob dieser Algorithmus anwendbar sein kann bzw. soll auf Frauen mit Proportionen außerhalb der Norm und ob das auschließen das Modell beeinträchtigt. Die Erste Frage lässt sich nur von einem Forscher in dieser Disziplin beantworten, die Zweite Frage hingegen wird im Fazit wieder betrachtet. Insofern werden die Ausreißer ersteinmal nicht entfernt.

Zur Frage ob es Abhängigkeiten zwischen den Features und dem Target gibt, lässt sich dazu eine Grundidee fassen, indem man verschiedene Features im zwei Dimensionalen gegeneinander aufträgt.3

Ebenfalls lässt sich mittels von Korrellationstabellen ebenfalls Korrellationen4 nachweisen, weswegen nun nur noch die geignete Methode zu Erstellung eines Prediction-Algorithmus benötigt wird.

3 Methoden

Es existieren zwei Kategorien in welche unsere Features einsortiert werden, insofern seien Algorithmen notwendig welche eine Klassifikation durchführen. Für kleinere Datensätze, wie hier vorliegend, ist *RandomForestClassifier* gut zu nutzen und da bereits aus Graph eine perfekte lineare separierung nicht möglich ist, ist *SVM (Support Vector Machine)* ebenfalls ein passender Algorithmus zur Modell erstellung.

3.1 Random Forest Classifier

Der Random Forest Classifier nutzt eine gewisse Anzahl an verschiedenen Entscheidungsbäumen mit dem Ziel die korrekte Entscheidung zu erfüllen, dies führt dazu, dass die Nachteile eines Entscheidungsbaumes verringert (wie z.B. *overfitting*)

3.2 Support Vector Machine

Ein SVM nutzt die Tatsache aus, dass selbst wenn eine Separierbarkeit z.B. im 2-Dimensionalen nicht machbar ist, diese in höheren Dimensionen durchaus möglich ist.

3.2.1 Kernel

- **rbf** Die gängigste Methode
- **linear** Dies diente mehr zur Überprüfung, ob dieses Problem nicht auch durch eine lineare Separierung teilbar ist.
  Wie aber hier zusehen, ist diese Lösung nicht optimal
- **poly**

4 Ergebnisse

Um das bestmöglichste Ergebnis zu erzielen, wurde ein Grid angewandt damit die bessere Anzahl an Entscheidungsbäumen bzw. des Wertes der Fehlklassifizierungsstrafe genutzt werden kann. Im folgenden werden die Modelle mitsamt ihrer Güte vorgestellt.

4.1 Random Forest Classifier

Für die Wahl der Anzahl an Bäumen ist laut der Gridanalyse5, 200 am passendsten, weitere Optimierungen seien zwar möglich, nicht aber zielführend, da höhere Genauigkeiten signifikant mehr Rechenzeit benötigen würden.

Zur Erstellung des Modells mittels des *Random Forest Classifiers* wurden die Daten in zwei Teile unterschieden, da 30% der Daten als Testdaten dienen sollen. Da leider nicht übermäßig viele Daten vorliegen, der Datensatz ist kleiner als tausend, sollte der Testdaten Anteil zumindest über 150 liegen. Mit diesen Eigenschaften ergab sich eine Accuracy von 0.9708 und eine Precision von 0.9725. Im Vergleich dazu besaß das Modell mit nur 10 Entscheidungsbäumen eine Accuracy von 0.9532 und eine Precision von 0.9630.
Veranschaulicht wird die Güte dieses Modells auch noch mittels einer Confusion Matrix6, welche die Fehlklassifikation von false positives bzw. false negatives zeigt.

4.2 Support Vector Machine

Für die passende Wahl an Parameter der Support Vector Machine gibt es zwei Parameter welche die größte Signifikanz besitzen, zum einen die Fehlklassifizierungsstrafe aber auch die passende Kernelmethode, ersichtlich in der Graphik7 ist die passende Fehlklassifizierungsstrafe 100. Für die passende Wahl der Kernelmethode ist ausprobieren meist die schnellste Methode

- **rbf** Die gängigste Methode sie ergab 0.9708 Accuracy sowie 0.9722 Precision und ebenfalls eine ähnlich confusion Matrix8
- **linear** Diese Methode ergab 0.9766 Accuracy sowie 0.9905 Precision und ebenfalls eine ähnlich confusion Matrix9.
- **poly** Diese Methode ergab 0.9649 Accuracy sowie 0.9550 Precision und ebenfalls eine ähnlich confusion Matrix**??**.

5 Diskussion

Beide Systeme besitzen zwar ähnliche Akkuratheit, dennoch besitzt das SVM Modell mit linearem Kernel die höchste Akkuratheit, sowie die höchste Precision. Stellt man die beiden Gegenüber11 wird eine kleine Überlegenheit des SVM Modells sichtbar.

6 Fazit

Generell erlaubt dieses Modell nun eine recht akkurate Einschätzung von Brustkrebs, mangels allzu umfangreicher Datenmenge sei dies aber nur mit Vorsicht zu genießen. Zusätzlich besitzt trotz der Außreiser

A Anhang: Erklärung

B Anhang: Grafiken

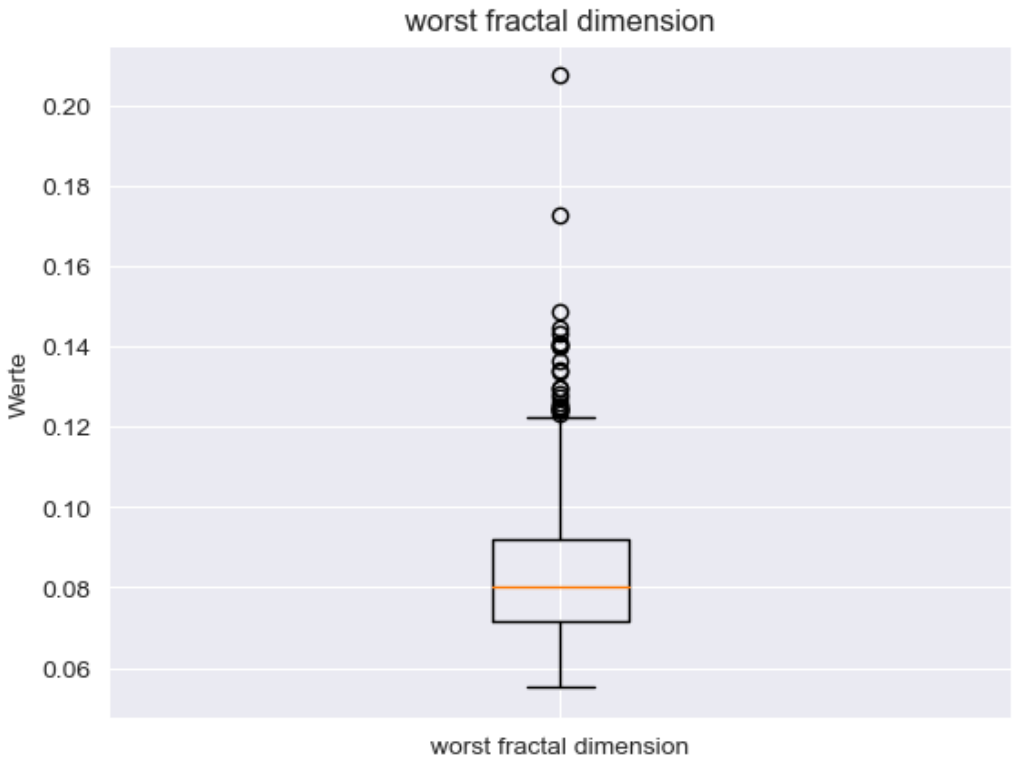

Abbildung 1: Boxplot der worst fractal Dimension

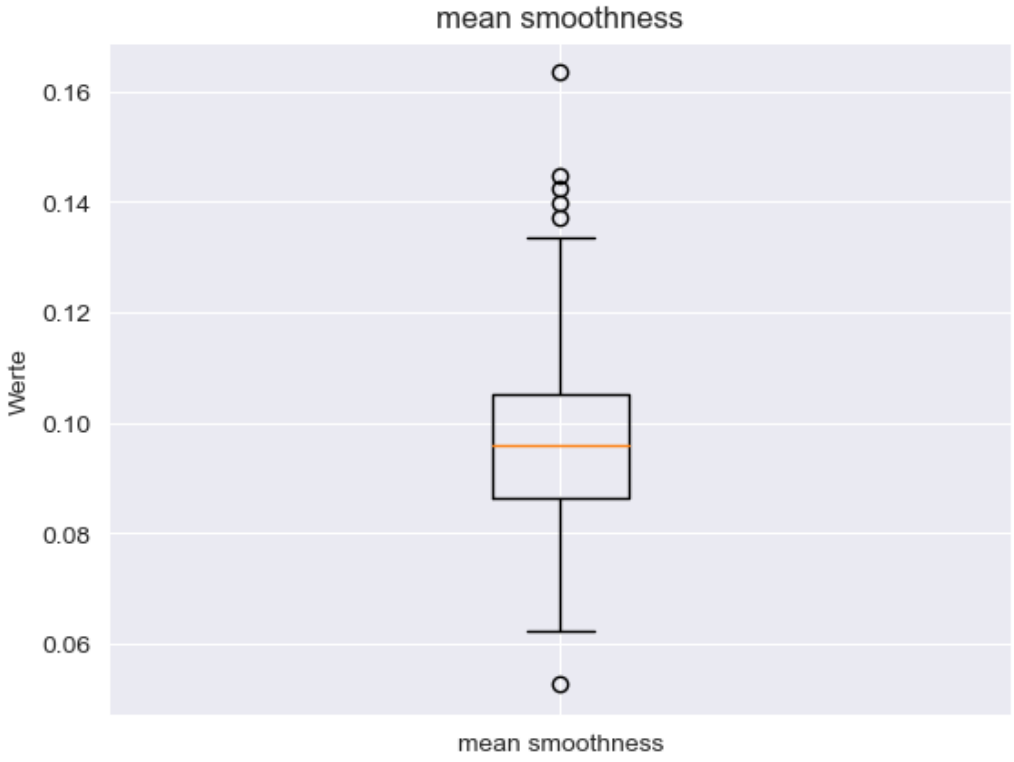

Abbildung 2: Boxplot der mean smoothness

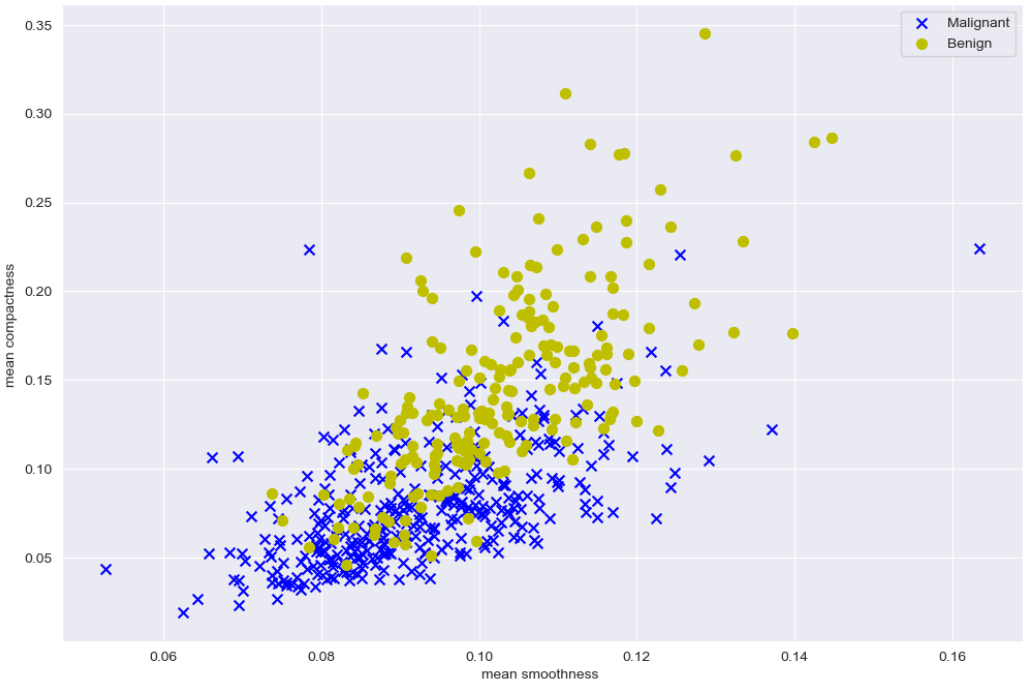

Abbildung 3: Mean smoothness und mean compactness

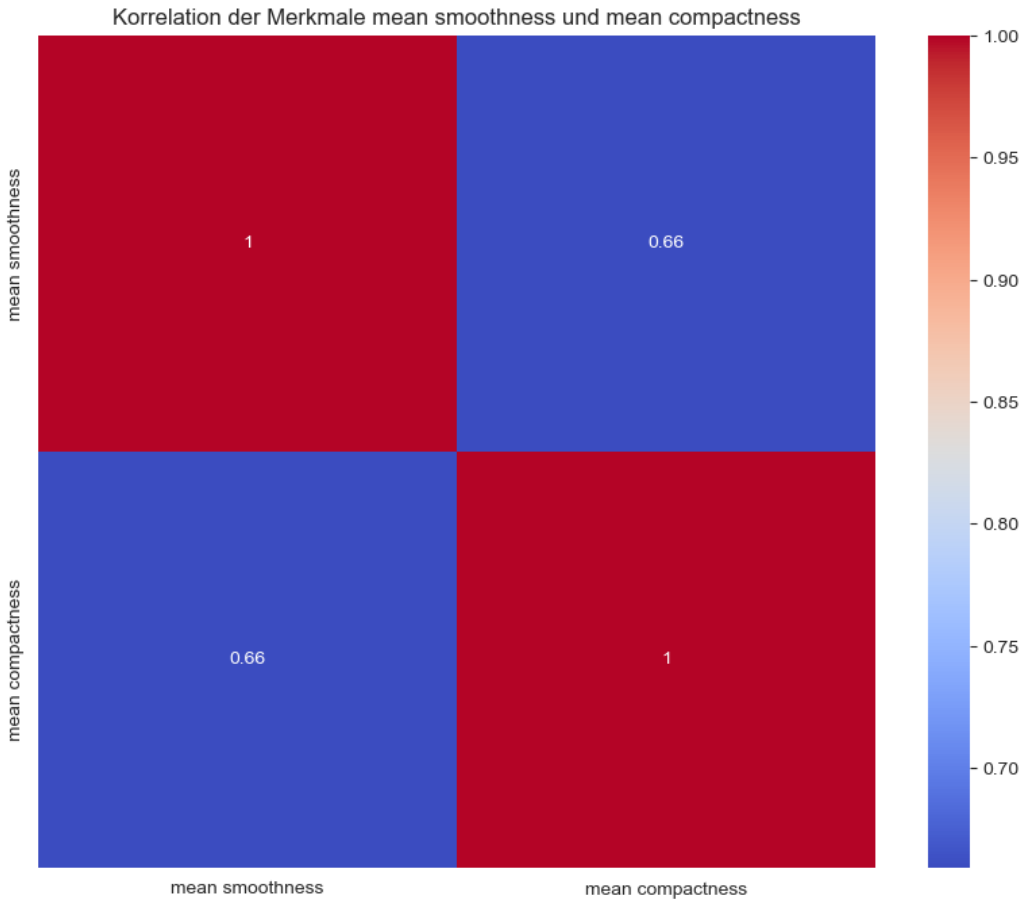

Abbildung 4: Korelation von

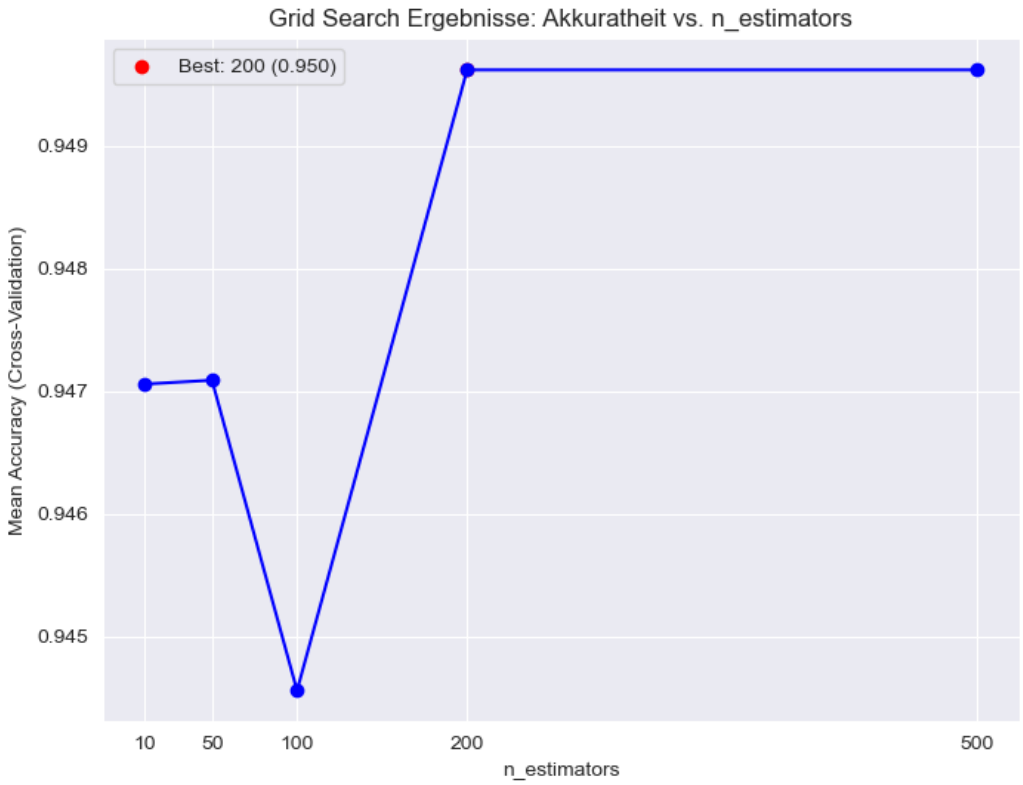

Abbildung 5: Jupyter Notebook als Anhang

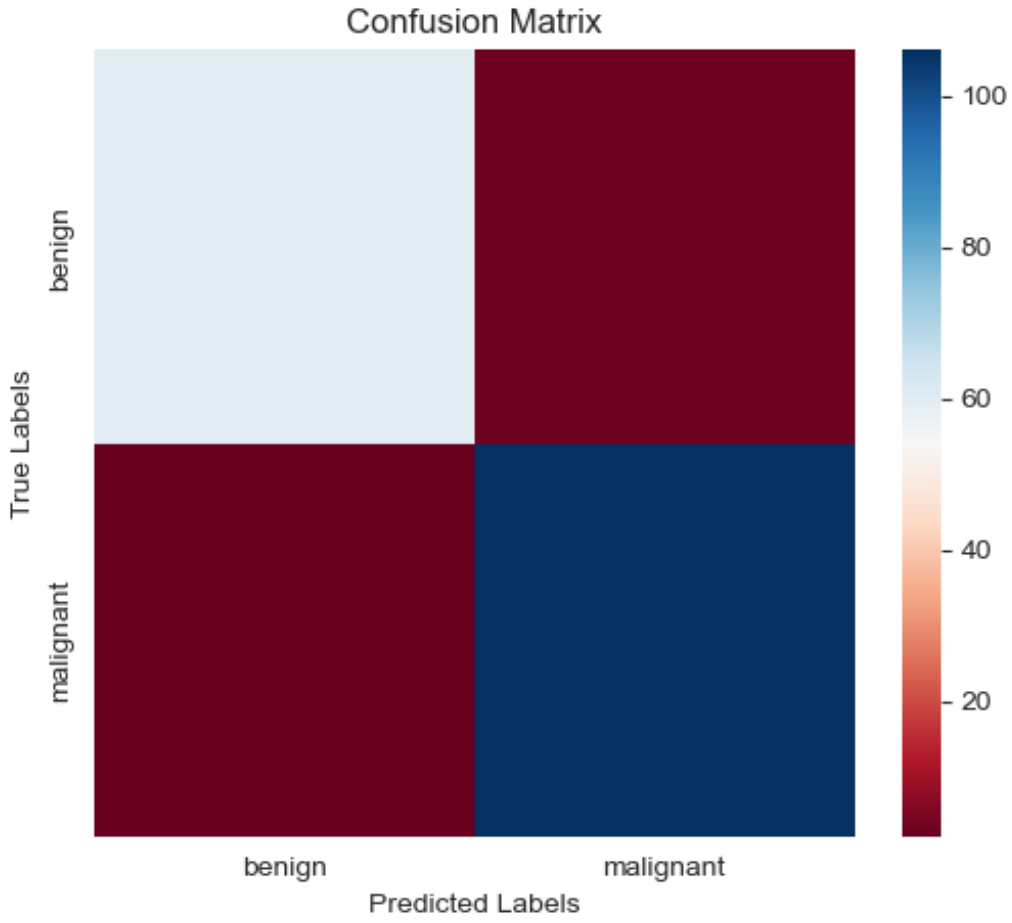

Abbildung 6: Jupyter Notebook als Anhang

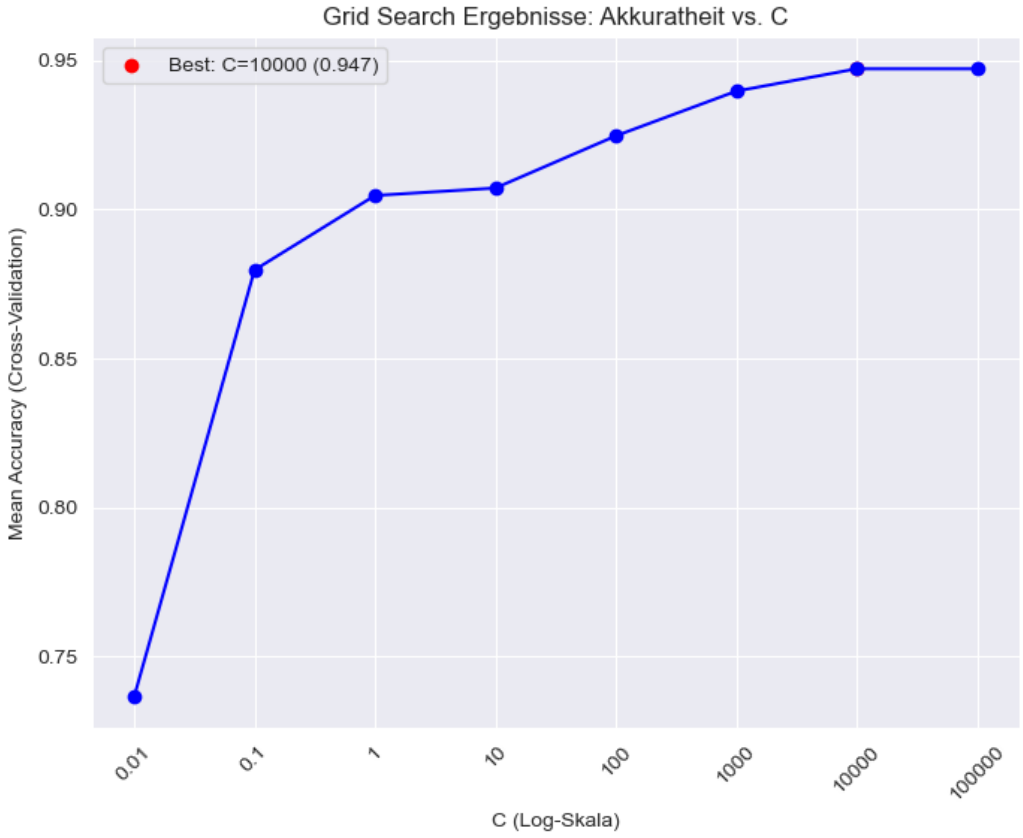

Abbildung 7: Jupyter Notebook als Anhang

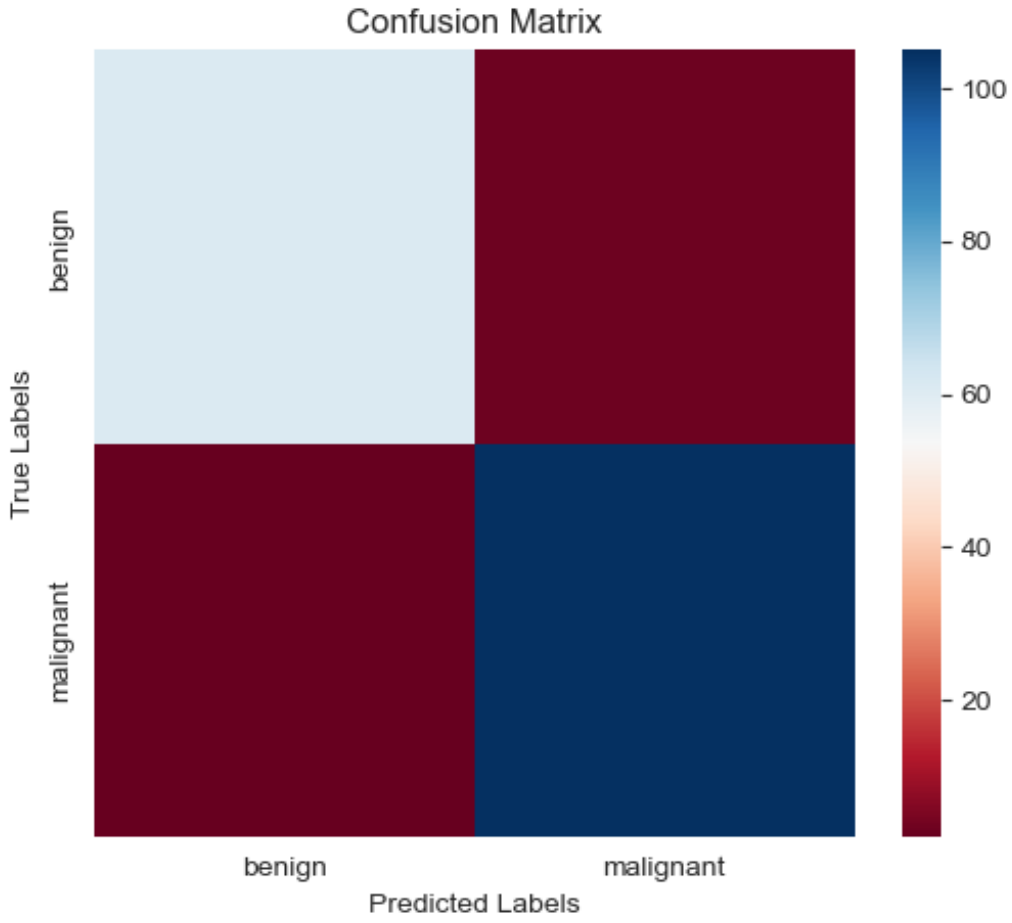

Abbildung 8: Jupyter Notebook als Anhang

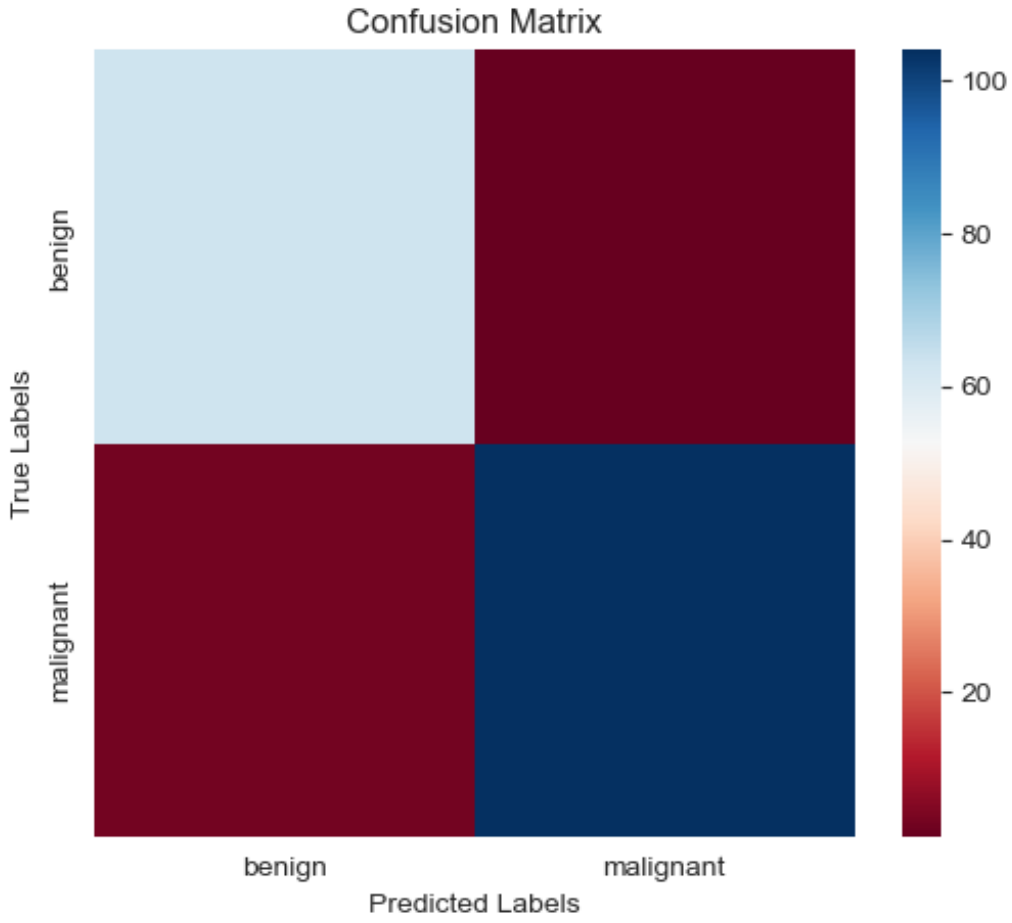

Abbildung 9: Jupyter Notebook als Anhang

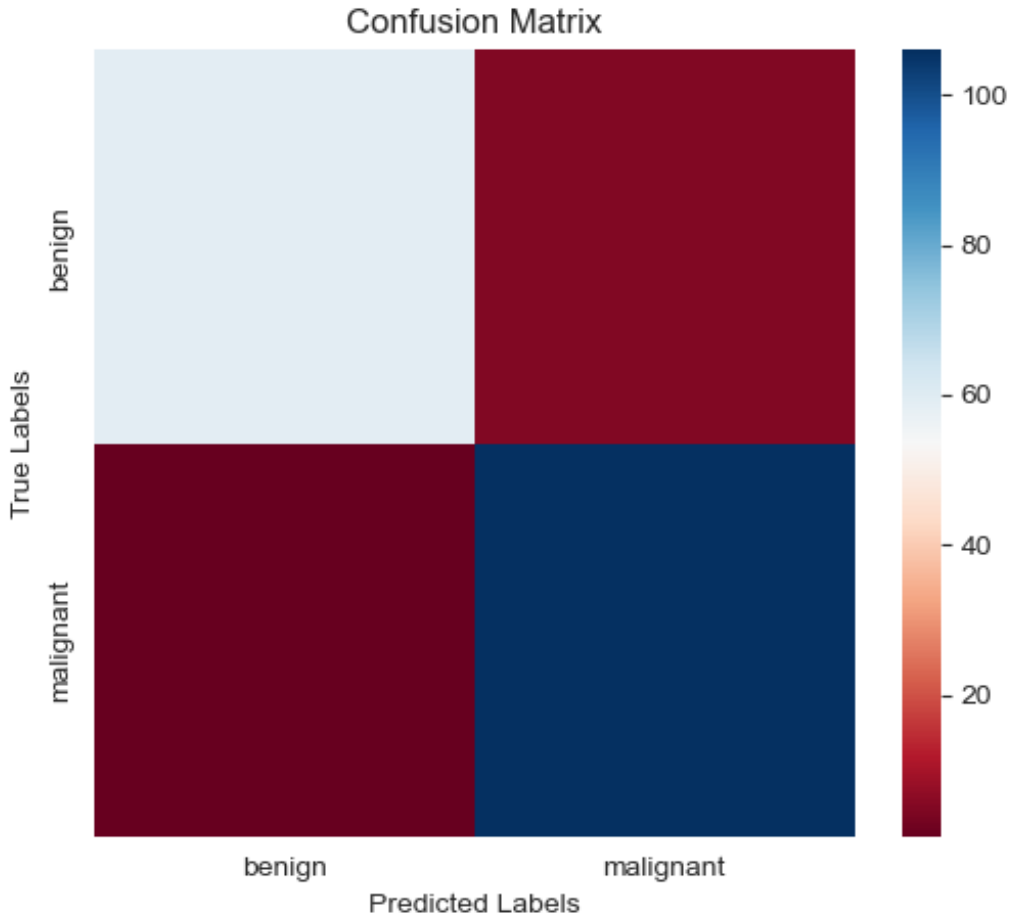

Abbildung 10: Jupyter Notebook als Anhang

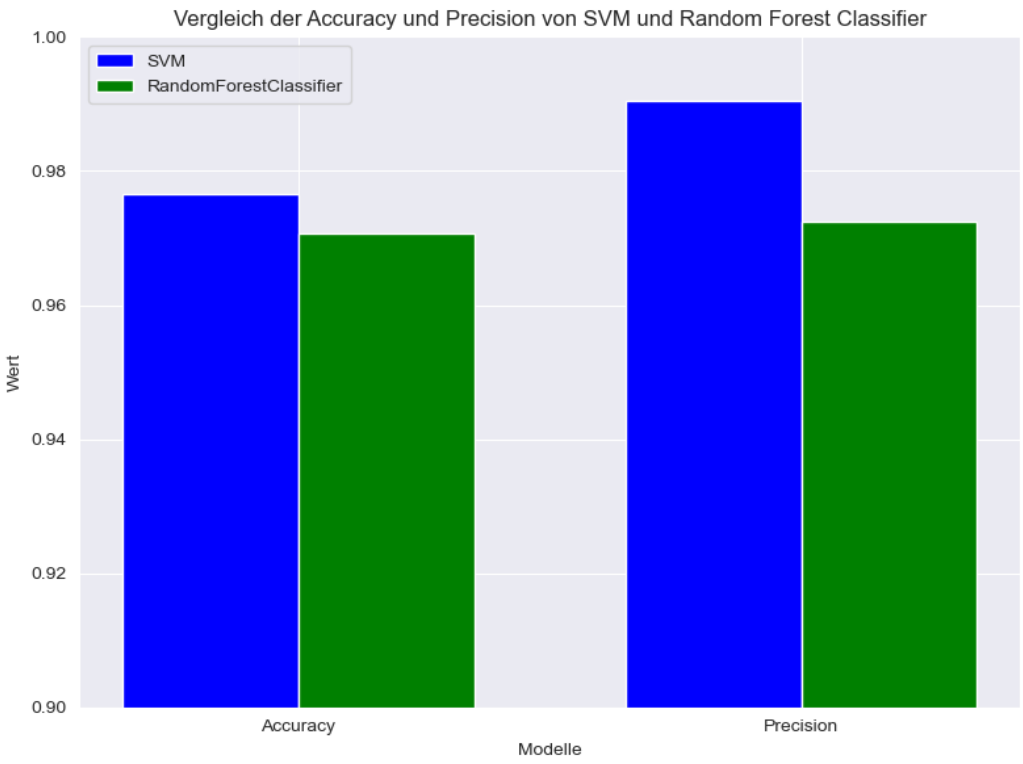

Abbildung 11: Vergleich der Accuracy und Precision von SVM und Random Forest Classifier